BEST AVAILABLE COPY

Freeform Search

Database:
US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Term:
L8 and (description near scheme)

Display: 100 **Documents in Display Format:** - **Starting with Number** 1

Generate: ○ **Hit List** ◉ **Hit Count** ○ **Side by Side** ○ **Image**

| Search | Clear | Interrupt |

Search History

DATE: Wednesday, June 29, 2005   Printable Copy   Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L9	L8 and (description near scheme)	9	L9
L8	L7 and semantic	14	L8
L7	L6 and scheme	16	L7
L6	structure and L5	21	L6
L5	L4 and description	23	L5
L4	L3 and fuzzy	23	L4
L3	L2 and (multimedia near content)	219	L3
L2	L1 and multimedia	454	L2
L1	mpeg-7	715	L1

END OF SEARCH HISTORY

☐ | Generate Collection | Print |

L9: Entry 9 of 9   File: USPT   May 13, 2003

US-PAT-NO: 6564263
DOCUMENT-IDENTIFIER: US 6564263 B1

TITLE: Multimedia content description framework

DATE-ISSUED: May 13, 2003

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Bergman; Lawrence David	Mt. Kisco	NY		
Kim; Michelle Yoonk Yung	Scarsdale	NY		
Li; Chung-Sheng	Ossining	NY		
Mohan; Rakesh	Stamford	CT		
Smith; John Richard	New Hyde Park	NY		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
International Business Machines Corporation	Armonk	NY			02

APPL-NO: 09/ 456031   [PALM]
DATE FILED: December 3, 1999

PARENT-CASE:
This application claims priority to U.S. Provisional Application Serial No.
60/110,902, filed on Dec. 4, 1998.

INT-CL: [07] G06 F 7/00, G06 F 15/00, G06 F 17/30, G06 F 15/16

US-CL-ISSUED: 709/231; 707/3, 707/101, 707/500.1, 707/104.1, 709/232
US-CL-CURRENT: 709/231; 707/101, 707/104.1, 707/3, 709/232, 715/500.1

FIELD-OF-SEARCH: 709/231, 709/232, 707/101, 707/104.1, 707/500.1, 725/53, 725/135,
725/136, 725/137

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

| Search Selected | Search ALL | Clear |

	PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
☐	6014671	January 2000	Castelli et al.	707/101

☐	6061689	May 2000	Chang et al.	707/103
☐	6181332	January 2001	Salahshour et al.	345/302
☐	6181817	January 2001	Zabith et al.	382/170
☐	6223183	April 2001	Smith et al.	707/102
☐	6232974	May 2001	Horvitz et al.	345/419
☐	6249423	May 2001	Hirata	707/104
☐	6282549	August 2001	Hoffert et al.	707/104
☐	6317795	November 2001	Malkin et al.	709/246
☐	6326965	December 2001	Castelli et al.	345/420
☐	6345279	February 2002	Li et al.	707/104
☐	6377995	April 2002	Agraharam et al.	709/231
☐	6411724	June 2002	Vaithiligam et al.	382/100

OTHER PUBLICATIONS

Chung-Sheng Li, Rakesh Mohan, John R. Smith, "Multimedia Content Description In The InfoPyramid", May 1998, IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP).*
John R. Smith, Rakesh Mohan, Chung-Sheng Li, "Scalable Multimedia Delivery for Pervasive Computing", Oct. 1999, ACM Multimedia.*
John R. Smith, Rakesh Mohan, Chung-Sheng Li, "Content-Based Transcoding of Images In the Internet", Oct. 1998, Proc. IEEE Proc. Int. Conf. Image Processing (ICIP), Chicago, Il,.*
Rakesh Mohan, John R. Smith, Chung-Sheng Li, "Adapting Multimedia Internet Content for Universal Access", Mar. 1999, IEEE Transactions on Multimedia, vol. 1, No. 1.

ART-UNIT: 2753

PRIMARY-EXAMINER: Dinh; Dung C.

ASSISTANT-EXAMINER: Johnson; Marlon

ATTY-AGENT-FIRM: Cameron; Douglas W.

ABSTRACT:

A framework is provided for describing multimedia content and a system in which a plurality of multimedia storage devices employing the content description methods of the present invention can interoperate. In accordance with one form of the present invention, the content description framework is a description scheme (DS) for describing streams or aggregations of multimedia objects, which may comprise audio, images, video, text, time series, and various other modalities. This description scheme can accommodate an essentially limitless number of descriptors in terms of features, semantics or metadata, and facilitate content-based search, index, and retrieval, among other capabilities, for both streamed or aggregated multimedia objects.

2 Claims, 19 Drawing figures

☐ ░░Generate Collection░░ | Print

L9: Entry 7 of 9   File: PGPB   May 16, 2002

PGPUB-DOCUMENT-NUMBER: 20020059584
PGPUB-FILING-TYPE: new
DOCUMENT-IDENTIFIER: US 20020059584 A1

TITLE: Audiovisual management system

PUBLICATION-DATE: May 16, 2002

INVENTOR-INFORMATION:

NAME	CITY	STATE	COUNTRY	RULE-47
Ferman, Ahmet Mufit	Vancouver	WA	US	
van Beek, Petrus J.L.	Vancouver	WA	US	
Errico, James H.	Portland	OR	US	

APPL-NO: 09/ 823377   [PALM]
DATE FILED: March 30, 2001

RELATED-US-APPL-DATA:
Application is a non-provisional-of-provisional application 60/232377, filed
September 14, 2000,

INT-CL: [07] H04 N 7/025, H04 N 7/10

US-CL-PUBLISHED: 725/34; 725/46
US-CL-CURRENT: 725/34; 725/46

REPRESENTATIVE-FIGURES: 2

ABSTRACT:

A system for structuring usage history for audiovisual materials.

☐   Generate Collection | Print |

L9: Entry 8 of 9   File: USPT   Jan 25, 2005

US-PAT-NO: 6847980
DOCUMENT-IDENTIFIER: US 6847980 B1

TITLE: Fundamental entity-relationship models for the generic audio visual data signal description

DATE-ISSUED: January 25, 2005

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Benitez; Ana B.	New York	NY	10027	
Jaimes; Alejandro	New York	NY	10021	
Chang; Shih-Fu	New York	NY	10027	
Smith; John R.	New York	NY	10025	
Li; Chung-sheng	Ossining	NY	10562	

APPL-NO: 09/ 607974   [PALM]
DATE FILED: June 30, 2000

PARENT-CASE:
CROSS REFERENCE TO RELATED APPLICATION The present application is based on U.S. provisional patent application No. 60/142,325, filed Jul. 3, 1999, from which priority is claimed.

INT-CL: [07] G06 F 17/30

US-CL-ISSUED: 707/104.1
US-CL-CURRENT: 707/104.1

FIELD-OF-SEARCH: 707/104.1, 707/500.1, 707/102, 707/6, 707/205, 345/328

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

| Search Selected | Search ALL | Clear |

	PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
☐	6223183	April 2001	Smith et al.	707/102
☐	6236395	May 2001	Sezan et al.	345/328
☐	6360234	March 2002	Jain et al.	707/500.1

ART-UNIT: 2171

PRIMARY-EXAMINER: Metjahic; Safet

ASSISTANT-EXAMINER: Alaubaidi; Haythim J

ATTY-AGENT-FIRM: Baker Botts LLP

ABSTRACT:

An invention for generating standard <u>description</u> records from <u>multimedia</u>
information. The invention utilizes fundamental entity-relation models for the
Generic AV DS that classify the entities, the entity attributes, and the
relationships in relevant types to describe visual data. It also involves
classification of entity attributes into syntactic and <u>semantic</u> attributes.
Syntactic attributes can be categorized into different levels: type/technique,
global distribution, local <u>structure,</u> and global composition. <u>Semantic</u> attributes
can be likewise discretely categorized: generic object, generic scene, specific
object, specific scene, abstract object, and abstract scene. The invention further
classifies entity relationships into syntactic<u>/semantic</u> categories. Syntactic
relationship categories include spatial, temporal, and visual categories. <u>Semantic</u>
relationship categories include lexical and predicative categories. Spatial and
temporal relationships can be topological or directional; visual relationships can
be global, local, or composition; lexical relationships can be synonymy, antonymy,
hyponymy/hypernymy, or meronymy/holonymy; and predicative relationships can be
actions (events) or states.

16 Claims, 12 Drawing figures

☐ | Generate Collection | | Print |

L9: Entry 9 of 9   File: USPT   May 13, 2003

US-PAT-NO: 6564263
DOCUMENT-IDENTIFIER: US 6564263 B1

TITLE: Multimedia content description framework

DATE-ISSUED: May 13, 2003

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Bergman; Lawrence David	Mt. Kisco	NY		
Kim; Michelle Yoonk Yung	Scarsdale	NY		
Li; Chung-Sheng	Ossining	NY		
Mohan; Rakesh	Stamford	CT		
Smith; John Richard	New Hyde Park	NY		

ASSIGNEE-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY	TYPE CODE
International Business Machines Corporation	Armonk	NY			02

APPL-NO: 09/ 456031   [PALM]
DATE FILED: December 3, 1999

PARENT-CASE:
This application claims priority to U.S. Provisional Application Serial No.
60/110,902, filed on Dec. 4, 1998.

INT-CL: [07] G06 F 7/00, G06 F 15/00, G06 F 17/30, G06 F 15/16

US-CL-ISSUED: 709/231; 707/3, 707/101, 707/500.1, 707/104.1, 709/232
US-CL-CURRENT: 709/231; 707/101, 707/104.1, 707/3, 709/232, 715/500.1

FIELD-OF-SEARCH: 709/231, 709/232, 707/101, 707/104.1, 707/500.1, 725/53, 725/135,
725/136, 725/137

PRIOR-ART-DISCLOSED:

U.S. PATENT DOCUMENTS

| Search Selected | | Search ALL | | Clear |

	PAT-NO	ISSUE-DATE	PATENTEE-NAME	US-CL
☐	6014671	January 2000	Castelli et al.	707/101

☐	6061689	May 2000	Chang et al.	707/103
☐	6181332	January 2001	Salahshour et al.	345/302
☐	6181817	January 2001	Zabith et al.	382/170
☐	6223183	April 2001	Smith et al.	707/102
☐	6232974	May 2001	Horvitz et al.	345/419
☐	6249423	May 2001	Hirata	707/104
☐	6282549	August 2001	Hoffert et al.	707/104
☐	6317795	November 2001	Malkin et al.	709/246
☐	6326965	December 2001	Castelli et al.	345/420
☐	6345279	February 2002	Li et al.	707/104
☐	6377995	April 2002	Agraharam et al.	709/231
☐	6411724	June 2002	Vaithiligam et al.	382/100

OTHER PUBLICATIONS

Chung-Sheng Li, Rakesh Mohan, John R. Smith, "Multimedia Content Description In The InfoPyramid", May 1998, IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP).*
John R. Smith, Rakesh Mohan, Chung-Sheng Li, "Scalable Multimedia Delivery for Pervasive Computing", Oct. 1999, ACM Multimedia.*
John R. Smith, Rakesh Mohan, Chung-Sheng Li, "Content-Based Transcoding of Images In the Internet", Oct. 1998, Proc. IEEE Proc. Int. Conf. Image Processing (ICIP), Chicago, Il,.*
Rakesh Mohan, John R. Smith, Chung-Sheng Li, "Adapting Multimedia Internet Content for Universal Access", Mar. 1999, IEEE Transactions on Multimedia, vol. 1, No. 1.

ART-UNIT: 2753

PRIMARY-EXAMINER: Dinh; Dung C.

ASSISTANT-EXAMINER: Johnson; Marlon

ATTY-AGENT-FIRM: Cameron; Douglas W.

ABSTRACT:

A framework is provided for describing multimedia content and a system in which a plurality of multimedia storage devices employing the content description methods of the present invention can interoperate. In accordance with one form of the present invention, the content description framework is a description scheme (DS) for describing streams or aggregations of multimedia objects, which may comprise audio, images, video, text, time series, and various other modalities. This description scheme can accommodate an essentially limitless number of descriptors in terms of features, semantics or metadata, and facilitate content-based search, index, and retrieval, among other capabilities, for both streamed or aggregated multimedia objects.

2 Claims, 19 Drawing figures

Previous Doc   Next Doc   Go to Doc#

L9: Entry 8 of 9   File: USPT   Jan 25, 2005

DOCUMENT-IDENTIFIER: US 6847980 B1
TITLE: Fundamental entity-relationship models for the generic audio visual data
signal description

Abstract Text (1):
An invention for generating standard description records from multimedia
information. The invention utilizes fundamental entity-relation models for the
Generic AV DS that classify the entities, the entity attributes, and the
relationships in relevant types to describe visual data. It also involves
classification of entity attributes into syntactic and semantic attributes.
Syntactic attributes can be categorized into different levels: type/technique,
global distribution, local structure, and global composition. Semantic attributes
can be likewise discretely categorized: generic object, generic scene, specific
object, specific scene, abstract object, and abstract scene. The invention further
classifies entity relationships into syntactic/semantic categories. Syntactic
relationship categories include spatial, temporal, and visual categories. Semantic
relationship categories include lexical and predicative categories. Spatial and
temporal relationships can be topological or directional; visual relationships can
be global, local, or composition; lexical relationships can be synonymy, antonymy,
hyponymy/hypernymy, or meronymy/holonymy; and predicative relationships can be
actions (events) or states.

Brief Summary Text (3):
The present invention relates to techniques for describing multimedia information,
and more specifically, to techniques which describe both video and image
information, or audio information, as well as to content of such information. The
techniques disclosed are for content-sensitive indexing and classification of
digital data signals (e.g., multimedia signals).

Brief Summary Text (4):
II. Description of the Related Art

Brief Summary Text (5):
With the maturation of the global Internet and the widespread employment of
regional networks and local networks, digital multimedia information has become
increasingly accessible to consumers and businesses. Accordingly, it has become
progressively more important to develop systems that process, filter, search and
organize digital multimedia information, so that useful information can be culled
from this growing mass of raw information.

Brief Summary Text (7):
Unfortunately, the same is not true for multimedia content, as no generally
recognized description of this material exists.

Brief Summary Text (10):
In both paradigms there are classification issues which are often overlooked,
particularly in the content-based retrieval community. The main difficulty in
appropriately indexing visual information can be summarized as follows: (I) there
is a large amount of information present in a single image (e.g., what to index?),
and (2) different levels of description are possible (e.g., how to index?).

Consider, for example, a portrait of a man wearing a suit. It would be possible to label the image with the terms "suit" or "man". The term "man", in turn, could carry information at multiple levels: conceptual (e.g., definition of man in the dictionary), physical (size, weight) and visual (hair color, clothing), among others. A category label, then, implies explicit (e.g., the person in the image is a man, not a woman), and implicit or undefined information (e.g., from that term alone it is not possible to know what the man is wearing).

Brief Summary Text (11):
In this regard, there have been past attempts to provide multimedia databases which permit users to search for pictures using characteristics such as color, texture and shape information of video objects embedded in the picture. However, at the closing of the 20th Century, it is not yet possible to perform a general search the Internet or most regional or local networks for multimedia content, as no broadly recognized description of this material exists. Moreover, the need to search for multimedia content is not limited to databases, but extends to other applications, such as digital broadcast television and multimedia telephony.

Brief Summary Text (12):
One industry wide attempt to develop such standard a multimedia description framework has been through the Motion Pictures Expert Group's ("MPEG") MPEG-7 standardization effort. Launched in October 1996, MPEG-7 aims to standardize content descriptions of multimedia data in order to facilitate content-focused applications like multimedia searching, filtering, browsing and summarization. A more complete description of the objectives of the MPEG-7 standard are contained in the International Organisation for Standardisation document ISO/IEC JTC1/SC29/WG11 N2460 (October 1998), the content of which is incorporated by reference herein.

Brief Summary Text (13):
The MPEG-7 standard has the objective of specifying a standard set of descriptors as well as structures (referred to as "description schemes") for the descriptors and their relationships to describe various types of multimedia information. MPEG-7 also proposes to standardize ways to define other descriptors as well as "description schemes" for the descriptors and their relationships. This description, i.e. the combination of descriptors and description schemes, shall be associated with the content itself, to allow fast and efficient searching and filtering for material of a user's interest. MPEG-7 also proposes to standardize a language to specify description schemes, i.e. a Description Definition Language ("DDL"), and the schemes for binary encoding the descriptions of multimedia content.

Brief Summary Text (14):
At the time of filing the instant application, MPEG is soliciting proposals for techniques which will optimally implement the necessary description schemes for future integration into the MPEG-7 standard. In order to provide such optimized description schemes, three different multimedia-application arrangements can be considered. These are the distributed processing scenario, the content-exchange scenario, and the format which permits the personalized viewing of multimedia content.

Brief Summary Text (15):
Regarding distributed processing, a description scheme must provide the ability to interchange descriptions of multimedia material independently of any platform, any vendor, and any application, which will enable the distributed processing of multimedia content. The standardization of interoperable content descriptions will mean that data from a variety of sources can be plugged into a variety of distributed applications, such as multimedia processors, editors, retrieval systems, filtering agents, etc. Some of these applications may be provided by third parties, generating a sub-industry of providers of multimedia tools that can work with the standardized descriptions of the multimedia data.

Brief Summary Text (16):
A user should be permitted to access various content providers' web sites to download content and associated indexing data, obtained by some low-level or high-level processing, and proceed to access several tool providers' web sites to download tools (e.g. Java applets) to manipulate the heterogeneous data descriptions in particular ways, according to the user's personal interests. An example of such a multimedia tool will be a video editor. A MPEG-7 compliant video editor will be able to manipulate and process video content from a variety of sources if the description associated with each video is MPEG-7 compliant. Each video may come with varying degrees of description detail, such as camera motion, scene cuts, annotations, and object segmentations.

Brief Summary Text (17):
A second scenario that will greatly benefit from an interoperable content description standard is the exchange of multimedia content among heterogeneous multimedia databases. MPEG-7 aims to provide the means to express, exchange, translate, and reuse existing descriptions of multimedia material.

Brief Summary Text (18):
Currently, TV broadcasters, Radio broadcasters, and other content providers manage and store an enormous amount of multimedia material. This material is currently described manually using textual information and proprietary databases. Without an interoperable content description, content users need to invest manpower to translate manually the descriptions used by each broadcaster into their own proprietary scheme. Interchange of multimedia content descriptions would be possible if all the content providers embraced the same scheme and content description schemes. This is one of the objectives of MPEG-7.

Brief Summary Text (19):
Finally, multimedia players and viewers that employ the description schemes must provide the users with innovative capabilities such as multiple views of the data configured by the user. The user should be able to change the display's configuration without requiring the data to be downloaded again in a different format from the content broadcaster.

Brief Summary Text (20):
The foregoing examples only hint at the possible uses for richly structured data delivered in a standardized way based on MPEG-7. Unfortunately, no prior art techniques available at present are able to generically satisfy the distributed processing, content-exchange, or personalized viewing scenarios. Specifically, the prior art fails to provide a technique for capturing content embedded in multimedia information based on either generic characteristics or semantic relationships, or to provide a technique for organizing such content. Accordingly, there exists a need in the art for efficient content description schemes for generic multimedia information.

Brief Summary Text (21):
During the MPEG Seoul Meeting (March 1999), a Generic Visual Description Scheme (Video Group, "Generic Visual Description Scheme for MPEG-7", ISO/IEC JTC1/SC29/WG11 MPEG99/N2694, Seoul, Korea, March 1999) was generated following some of the recommendations from the DS1 (still images), DS3++ (multimedia), DS4 (application), and, especially, DS2 (video) teams of the MPEG-7 Evaluation AHG (Lancaster, U.K., February 1999) (AHG on MPEG-7 Evaluation Logistics, "Report of the Ad-hoc Group on MPEG-7 Evaluation Logistics", ISO/IEC JTC1/SC29/WG11 MPEG99/N4524, Seoul, Korea, March 1999). The Generic Visual DS has evolved in the AHG on Description Schemes to the Generic Audio Visual Description Scheme ("AV DS") (AHG on Description Scheme, "Generic Audio Visual Description Scheme for MPEG-7 (V0.3)", ISO/IEC JTC1/SC29/WG11 MPEG99/M4677, Vancouver, Canada, July 1999). The Generic AV DS describes the visual content of video sequences or images and,

partially, the content of audio sequences; it does not address <u>multimedia</u> or archive content.

<u>Brief Summary Text</u> (22):
The basic components of the Generic AV DS are the syntactic <u>structure</u> DS, the <u>semantic structure</u> DS, the syntactic-<u>semantic</u> links DS, and the analytic/synthetic model DS. The syntactic <u>structure</u> DS is composed of region trees, segment trees, and segment/region relation graphs. Similarly, the <u>semantic structure</u> DS is composed of object trees, event trees, and object/event relation graphs. The syntactic-<u>semantic</u> links DS provide a mechanism to link the syntactic elements (regions, segments, and segment/region relations) with the <u>semantic</u> elements (objects, events, and event/object relations), and vice versa. The analytic/synthetic model DS specifies the projection/registration/conceptual correspondence between the syntactic and the <u>semantic structure</u>. The <u>semantic</u> and syntactic elements, which we will refer to as content elements in general, have associated attributes. For example, a region is described by color/texture, shape, 2-D geometry, motion, and deformation descriptors. An object is described by type, object-behavior, and <u>semantic</u> annotation DSs.

<u>Brief Summary Text</u> (23):
We have identified possible shortcomings in the current specification of the Generic AV DS. The Generic AV DS includes content elements and entity-relation graphs. The content elements have associated features, and the entity-relation graphs describe general relationships among the content elements. This follows the Entity-Relationship (ER) modeling technique (P. P-S. Chen, "The Entity-Relation Model--Toward a Unified View of Data", ACM Transactions on Database Systems, Vol. 1, No. 1, pp. 9-36, March 1976). The current specification of these elements in the Generic AV DS, however, is too generic to become a useful and powerful tool to describe audio-visual content. The Generic AV DS also includes hierarchies and links between the hierarchies, which is typical of physical hierarchical models. Consequently, the Generic AV DS is a mixture of different conceptual and physical models. Other limitations of this DS may be the rigid separation of the <u>semantic</u> and the syntactic <u>structures</u> and the lack of explicit and unified definitions of its content elements.

<u>Brief Summary Text</u> (24):
The Generic AV DS describes images, video sequences, and, partially, audio sequences following the classical approach for book content <u>descriptions</u>: (1) definition of the physical or syntactic <u>structure</u> of the document; the Table of Contents; (2) definition of the <u>semantic structure,</u> the Index; and (3) definition of the locations where <u>semantic</u> notions appear. It consists of (1) syntactic <u>structure</u> DS; (2) <u>semantic structure</u> DS; (3) syntactic-<u>semantic</u> links DS; (4) analytic/synthetic model DS; (5) visualization DS; (6) meta information DS; and (7) media information DS.

<u>Brief Summary Text</u> (25):
The syntactic DS is used to specify physical <u>structures</u> and the signal properties of an image or a video sequence defining the table of contents of the document. It consists of (1) segment DS; (2) region DS; and (3) segment/region relation graph DS. The segment DS may be used to define trees of segments that specify the linear temporal <u>structure</u> of the video program. Segments are a group of continuous frames in a video sequence with associated features: time DS, meta information DS, media information DS. A special type of segment, a shot, includes editing effect DS, key frame DS, mosaic DS, and camera motion DS. Similarly, the region DS may be used to define a tree of regions. A region is defined as group of connected pixels in a video sequence of an image with associated features: geometry DS, color/texture DS, motion DS, deformation DS, media information DS, and meta information DS. The segment/region relation graph DS specifies general relationships among segments and regions, e.g. spatial relationships such as "To The Left Of"; temporal relationships such as "Sequential To"; and <u>semantic</u> relationships such as "Consist

of".

Brief Summary Text (26):
The semantic DS is used to specify semantic features of an image or a video
sequence in terms of semantic objects and events. It can be viewed as a set of
indexes. It consists of (1) event DS; (2) object DS; and (3) event/object relation
graph DS. The event DS may be used to form trees of events that define a semantic
index table for the segments in the segment DS. Events contain an annotation DS.
Similarly, the object DS may be used to form trees of objects that define a
semantic index table for the objects in the object DS. The event/object relation
graph DS specifies general relationships among events and objects.

Brief Summary Text (27):
The syntactic-semantic links DS are bi-directional between the syntactic elements
(segments, regions, or segment/region relations) and the semantic elements (events,
objects, or event/object relations). The analytic/synthetic model DS specifies the
projection/registration/conceptual correspondence between syntactic and semantic
structure DSs. The media and meta information DS contains descriptors of the
storage media and the author-generated information, respectively. The visualization
DS contains a set of view DS to enable efficient visualization of a video program.
It includes the following views: multi-resolution space-frequency thumbnail, key-
frame, highlight, event, and alternate views. Each one of these views is
independently defined.

Brief Summary Text (29):
The Generic AV DS includes content elements (i.e. regions, objects, segments, and
events), with associated features. It also includes entity-relation graphs to
describe general relationships among content elements following the entity-
relationship model. A drawback of the current DS is that the features and the
relationships among elements can have a broad range of values, which reduces their
usefulness and expressive power. A clear example is the semantic annotation feature
in the object element. The value of the semantic annotation could be a generic
("Man"), a specific ("John Doe"), or an abstract ("Happiness") concept.

Brief Summary Text (30):
The initial goal of the development leading to the present invention was to define
explicit entity-relationship structures for the Generic AV DS to address this
drawback. The explicit entity-relationship structures would categorize the
attributes and the relationships into relevant classes. During this process,
especially during the generation of concrete examples (see the baseball example
shown in FIGS. 6-9), we became aware of other shortcomings of the current Generic
AV DS, this time, related to the DS's global design. We shall present these in this
section. In this application, we propose complete fundamental entity-relationship
models that try to address these issues.

Brief Summary Text (31):
First, the full specification of the Generic DS could be represented using an
entity-relationship model. As an example, the entity-relation models provided in
FIGS. 7-9 for the baseball example in FIG. 6, include the functionality addressed
by most of the components of the Generic AV DS (e.g. the event DS, the segment DS,
the object DS, the region DS, the syntactic-semantic links DS, the segment/region
relation graph DS, and the event/object relation graph DS) and more. The entity-
relationship (E-R) model is a popular high-level conceptual data model, which is
independent of the actual implementation as hierarchical, relational, or object-
oriented models, among others. The current version of the Generic DS seems to be a
mix of multiple conceptual and implementation data models: the entity-relationship
model (e.g. segment/region relation graph), the hierarchical model (e.g. region DS,
object DS, and syntactic-semantic links DS), and the object-oriented model (e.g.
segment DS, visual segment DS, and audio segment DS).

Brief Summary Text (32):
Second, the separation between syntax and semantics in the current Generic DS is
too rigid. For the example in FIG. 6, we have separated the descriptions of the
Batting Event and the Batting Segment (see FIG. 7), as the current Generic AV DS
proposes. In this case, however, it would have been more convenient to merge both
elements into a unique Batting Event with semantic and syntactic features. Many
groups working on video indexing have advocated the separation of the syntactic
structures (Table of Contents: segments and shots) and the semantic structures
(Semantic Indexes: events). In describing images or animated objects in video
sequences, however, the value of separating these structures is less clear. "Real
objects" are usually described by their semantic features (e.g. semantic class--
person, cat, etc.) as well as by their syntactic features (e.g. color, texture, and
motion). The current Generic AV DS separates the definition of "real objects" in
the region and the object DSs, which may cause inefficient handling of the
descriptions.

Brief Summary Text (33):
Finally, the content elements, especially the object and the event, lack explicit
and unified definitions in the Generic DS. For example, the current Generic DS
defines an object as having some semantic meaning and containing other objects.
Although objects are defined in the object DS, event/object relation graphs can
describe general relationships among objects and events. Furthermore, objects are
linked to corresponding regions in the syntactic DS by the syntactic-semantic links
DS. Therefore, the object has a distributed definition across many components of
the Generic Visual DS, which is less than clear. The definition of an event is very
similar and as vague.

Brief Summary Text (35):
The Entity-Relationship (E-R) model first presented in P. P-S. Chen, "The Entity-
Relation Model--Toward a Unified View of Data", ACM Transactions on Database
Systems, Vol. 1, No. 1, pp. 9-36, March 1976 describes data in terms of entities
and their relationships. Both entities and relationships can be described by
attributes. The basic components of the entity-relationship model are shown in FIG.
1. The entity, the entity attribute, the relationship, and the relationship
attribute correspond very closely to the noun (e.g. a boy and an apple), the
adjective (e.g. young), the verb (e.g. eats), and the verb complement (e.g.
slowly), which are essential components for describing general data. "A young boy
eats an apple slowly", which could be the description of a video shot, is
represented using an entity-relationship model in FIG. 2. This modeling technique
has been used to model the contents of pictures and their features for image
retrieval.

Brief Summary Text (38):
An object of the present invention is to provide content description schemes for
generic multimedia information.

Brief Summary Text (39):
Another object of the present invention is to provide techniques for implementing
standardized multimedia content description schemes.

Brief Summary Text (40):
A further object of the present invention is to provide an apparatus which permits
users to perform enhanced content-sensitive general searches on the Internet or
regional or local networks for multimedia content.

Brief Summary Text (41):
Still another object of the present invention is to provide systems and techniques
for capturing content embedded in multimedia information based on either generic
characteristics or semantic relationships.

Brief Summary Text (42):
Still a further object of the present invention is to provide a technique for
organizing content embedded in multimedia information based on distinction of
entity attributes into syntactic and semantic. Syntactic attributes can be
categorized into different levels: type/technique, global distribution, local
structure, and global composition. Semantic attributes can be categorized into
different levels: generic object, generic scene, specific object, specific scene,
abstract object, and abstract scene.

Brief Summary Text (43):
Yet a further object of the present invention is classification of entity
relationships into syntactic and semantic categories. Syntactic relationships can
be categorized into spatial, temporal, and audio categories. Semantic relationships
can be categorized into lexical and predicative categories. Spatial and temporal
relationships can be topological or directional; audio relationships can be global,
local, or composition; lexical relationships can be synonymy, antonymy,
hyponymy/hypernymy, or meronymy/holonymy; and predicative relationships can be
actions (events) or states.

Brief Summary Text (46):
This work is based on the conceptual framework for indexing visual information
presented in A. Jaimes and S.-F. Chang, "A Conceptual Framework for Indexing Visual
Information at Multiple Levels", Submitted to Internet Imaging 2000, which has been
adapted and extended for the Generic AV DS. The work in other references (e.g., S.
Paek, A. B. Benitez, S.-F. Chang, C.-S. Li, J. R. Smith, L. D. Bergman, A. Puri, C.
Swain, and J. Ostermann, "Proposal for MPEG-7 image description scheme", Proposal
to ISO/IEC JTC1/SC29/WG11 MPEG99/P480, Lancaster, U.K., February 1999) is relevant
because it separates the description of the content elements (objects) and the
specification of relationships among the content elements (with entity-relation
graphs and hierarchies, a particular case of entity-relation graph). By doing so,
it is clearly specifying an E-R Model.

Brief Summary Text (47):
We focus on the problem of multiple levels of description for indexing visual
information. We present a novel conceptual framework, which unifies concepts from
the literature in diverse fields such as cognitive psychology, library sciences,
art, and the more recent content-based retrieval. We make distinctions between
visual and non-visual information and provide the appropriate structures. The ten-
level visual structure presented provides a systematic way of indexing images based
on syntax (e.g., color, texture, etc.) and semantics (e.g., objects, events, etc.),
and includes distinctions between general concept and visual concept. We define
different types of relations (e.g., syntactic, semantic) at different levels of the
visual structure, and also use a semantic information table to summarize important
aspects related to an image (e.g., that appear in the non-visual structure).

Brief Summary Text (48):
Our structures place state-of-the art content-based retrieval techniques in
perspective, relating them to real user-needs and research in other fields. Using
structures such as the ones presented, is beneficial not only in terms of
understanding the users and their interests, but also in characterizing the
content-based retrieval problem according to the levels of descriptions used to
access visual information.

Brief Summary Text (49):
The present invention proposes to index the attributes of the content elements
based on the ten-level conceptual structure presented in A. Jaimes and S.-F. Chang,
"A Conceptual Framework for Indexing Visual Information at Multiple Levels",
Submitted to Internet Imaging 2000, which distinguishes the attributes based on
syntax (e.g. color and texture) and semantics (e.g. semantic annotations) as shown
in FIG. 3. The first four levels of the visual structure refer to syntax, and the

remaining six refer to <u>semantics</u>. The syntax levels are type/technique, global distribution, local <u>structure,</u> and global composition. The <u>semantic</u> levels are generic object, generic scene, specific object, specific scene, abstract object, and abstract scene.

<u>Brief Summary Text</u> (50):
We also propose explicit types of relationships among content elements in the entity-relation graphs of the Generic AV DS. We distinguish between syntactic and <u>semantic</u> relationships as shown in FIG. 4. Syntactic relationships are divided into spatial, temporal, and visual. Spatial and temporal attributes are classified into topological and directional classes. Syntactic-attribute relationships can be further indexed into global, local, and composition. <u>Semantic</u> relationships are divided into lexical and predicative. Lexical relationships are classified into synonymy, antonymy, hyponymy/hypernymy, and meronymy/holonymy. Predicative relationships can be further indexed into action and event.

<u>Brief Summary Text</u> (51):
In term of types of content elements, we propose to classify them into syntactic and <u>semantic</u> elements. Syntactic elements can be divided into region, animated-regions, and segment elements; <u>semantic</u> elements can be indexed in object, animated-object, and event elements. We provide explicit and unified definitions of these elements that are represented in the proposed fundamental models in term of their attributes and the relationships with other elements. Inheritance relationships among some of these elements are also specified.

<u>Drawing Description Text</u> (1):
BRIEF <u>DESCRIPTION</u> OF THE DRAWINGS

<u>Drawing Description Text</u> (4):
FIG. 3 represents the indexing visual <u>structure</u> by a pyramid;

<u>Drawing Description Text</u> (5):
FIG. 4 shows relationships as proposed at different levels of the visual <u>structure;</u>

<u>Drawing Description Text</u> (8):
FIG. 7 is a conceptual <u>description</u> of the Batting Event for the Baseball batting event image displayed in FIG. 6;

<u>Drawing Description Text</u> (9):
FIG. 8 is a conceptual <u>description</u> of the Hit and the Throw Events for the Batting Event of FIG. 6;

<u>Drawing Description Text</u> (10):
FIG. 9 is a conceptual <u>description</u> of the Field Object for the Batting Event of FIG. 6;

<u>Drawing Description Text</u> (13):
FIG. 12 illustrates relationships at different levels of the audio <u>structure</u>. Elements within the syntactic levels are related according to syntactic relationships. Elements within the <u>semantic</u> levels are related according to syntactic and <u>semantics</u> relationships.

<u>Detailed Description Text</u> (1):
<u>DESCRIPTION</u> OF THE PREFERRED EMBODIMENTS

<u>Detailed Description Text</u> (2):
We choose the modeling technique used herein because entity-relationship models are the most widely used conceptual models. They provide a high degree of abstraction and are hardware and software independent. There exits specific procedures to

transform these models into physical models for implementation, which are hardware and software dependent. Examples of physical models are the hierarchical model, the relational model, and the object-oriented model. The E-R conceptual framework in the context of MPEG-7 is discussed in J. R. Smith and C.-S. Li, ""An E-R Conceptual Modeling Framework for MPEG-7"", Contribution to ISO/IEC JTC1/SC29/WG11 MPEG99, Vancouver, Canada, July 1999.

Detailed Description Text (3):
As shown in FIG. 5, we make the distinction between syntax and semantics for attributes (or MPEG-7 descriptors), relationships, and content elements. Syntax refers to the way the content elements are arranged without considering the meaning of such arrangements. Semantics, on the other hand, deals with the meaning of those elements and of their arrangements. As will be discussed in the remainder of the section, syntactic and semantic attributes can refer to several levels (the syntactic levels are type, global distribution, local structure, and global composition; the semantic levels are generic object/scene, specific object/scene, and abstract object/scene; see FIG. 3. Similarly, syntactic and semantic relationships can be further divided into sub-types referring to different levels (syntactic relationships are categorized into spatial, temporal, and visual relationships at generic and specific levels; semantic relationships are categorized into lexical and predicative; see FIG. 4. We provide compact and clear definitions of the syntactic and semantic elements based on their associated types of attributes and relationships with other elements. An important difference with the Generic AV DS, however, is that our semantic elements include not only semantic attributes but also syntactic attributes. Therefore, if an application would rather not distinguish between syntactic and semantic elements, it can do so by implementing all the elements as semantic elements.

Detailed Description Text (4):
To clarify the explanation of the fundamental entity-relationships models, we will use the examples in FIGS. 6-9, FIG. 6 shows a video shot of a baseball game representing as a Batting Event and a Batting Segment (segment and event as defined in the Generic AV DS). FIG. 7 includes a possible description of the Batting Event as composed of a Field Object, a Hit Event, a Throw Event, a temporal relationship "Before" between the Throw and the Hit Events, and some visual attributes. FIG. 8 presents descriptions of the Throw and the Hit Events and relationships among them. The Throw Event is the action that the Pitcher Object executes over a Ball Object towards the Batter Object, "Throws". We provide some semantic attributes for the Pitcher Object. The Hit Event is the action that the Batter Object executes over the same Ball Object, "Hit". FIG. 9 shows the decomposition of the Field Object into three different regions, one of which is related to the Pitcher Object by the spatial relationships "On top of". Some visual attributes for one of these regions are provided.

Detailed Description Text (6):
We propose a ten-level conceptual structure to index the visual content elements (e.g. regions, entire images, and events) in image and video descriptions. This structure is valid only for the information explicitly depicted in the actual image or the video sequence (e.g., the price of a painting would not be part of visual content).

Detailed Description Text (7):
The proposed visual structure contains ten levels: the first four refer to syntax, and the remaining six refer to semantics. An overview of the visual structure is given in FIG. 3. The lower the level is in the pyramid, the more knowledge and information is required to perform indexing. The width of each level is an indication of the amount of knowledge required there. The indexing cost of an attribute can be included as a sub-attribute of the attribute. The syntax levels are type/technique, global distribution, local structure, and global composition. The semantic levels are generic object, generic scene, specific object, specific

scene, abstract object, and abstract scene. While some of these divisions may not be strict, they should be considered because they have a direct impact in understanding what the user is searching for and how he tries to find it in a database. They also emphasize the limitations of different indexing techniques (manual and automatic) in terms of the knowledge required.

Detailed Description Text (8):
In FIG. 3, the indexing visual_structure is represented by a pyramid. It is clear that the lower the level in the pyramid, the more knowledge and information is required to perform the indexing there. The width of each level is an indication of the amount of knowledge required--for example, more information is needed to name specific objects in the same scene.

Detailed Description Text (9):
In FIG. 5, the syntactic attribute (syntactic Ds) includes an enumerated attribute, level, whose value is its corresponding syntactic level in the visual structure (FIG. 3)--i.e. type, global distribution, local structure, or global composition-- or "not specified". The semantic attributes also include an enumerated attribute, level, whose value is its corresponding semantic level in the semantic structure (FIG. 3)--i.e. generic object, generic scene, specific object, specific scene, abstract object, and abstract scene--or "not specified". Another possibility of modeling the different types of syntactic and semantic attributes would be to subclass the syntactic and the semantic attribute elements to create type, global distribution, local structure, and global composition syntactic attributes; or generic object, generic scene, specific object, specific scene, abstract object, abstract scene attributes (some of these types do not apply for all object, animated object, and event), respectively.

Detailed Description Text (10):
Each level of the visual_structure is explained below. A discussion of the relationships between levels appears thereafter. Based on this visual structure and the relationships between levels, we define types of content elements in the following section.

Detailed Description Text (12):
At the most basic level, we are interested in the general visual characteristics of the image or the video sequence. Descriptions of the type of image or video sequence or the technique used to produce it are very general, but prove to be of great importance when organizing a visual database. Images, for example, may be placed in categories such as painting, black and white (b&w), color photograph, and drawing. Related classification schemes at this level have been done automatically in WebSEEk. The type for the example in FIG. 6 is color video sequence.

Detailed Description Text (16):
Local Structure

Detailed Description Text (17):
In processing the information of an image or video sequence, we perform different levels of grouping. In contrast to Global Structure, which does not provide any information about the individual parts of the image or the video sequence, the Local Structure level is concerned with the extraction and characterization of the components. At the most basic level, those components result from low-level processing and include elements such as the Dot, Line, Tone, Color, and Texture. As an example, a binary shape mask describes the Batting Segment in FIG. 6 (see FIG. 7). Other examples of local structure attributes are temporal/spatial position (e.g. start time and centroid), local color (e.g. M.times.N Layout), local motion, local deformation, local shape/2D geometry (e.g. bounding box).

Detailed Description Text (20):
At this level, we focus on the specific arrangement or composition of the basic

elements given by the local <u>structure</u>. In other words, we analyze the image as a whole, but only use the basic elements described in the previous level (e.g. line and circle) for the analysis. Global Composition refers to the arrangement or spatial layout of elements in the image. Traditional analysis in art describes composition concepts such as balance, symmetry, center of interest (center of attention or focus), leading line, and viewing angle. At this level, however, there is no knowledge of specific objects; only basic elements (e.g. dot, line, and circle.) or groups of basic elements are considered. The 2D geometry of the Sand 1 Region in FIG. 6 is a global composition attribute (see FIG. 9).

<u>Detailed Description Text</u> (21):
Generic Objects Up to the previous level, no world knowledge is required to perform indexing, so automatic techniques can be used to extract relevant information on these levels. Several studies, however, have demonstrated that humans mainly use higher level attributes to describe, classify and search for visual material. C. Jorgensen, "Image Attributes in Describing Tasks: an Investigation", Information Processing & Management, 34, (2/3), pp. 161-174, 1998. C. Jorgensen, "Retrieving the Unretrievable: Art, Aesthetics, and Emotion in Image Retrieval Systems", SPIE Conference in Human Vision and Electronic Imaging, IS&T/SPIE99, Vol. 3644, San Jose, Calif., January 1999. Objects are of particular interest, but they can also be placed in categories at different levels--an apple can be classified as a Macintosh apple, as an apple, or as a fruit. When referring to Generic Objects, we are interested in the basic level categories: the most general level of object <u>description,</u> which can be recognized with everyday knowledge. For the Pitcher Object in FIG. 6, a generic-object attribute could be the annotation "Man" (see FIG. 8).

<u>Detailed Description Text</u> (25):
In contrast to the previous level, Specific Objects refer to identified and named objects. Specific knowledge of the objects in the image or the video sequence is required, and such knowledge is usually objective since it relies on known facts. Examples include individual persons (e.g., the <u>semantic</u> annotation "Peter Who, Player #3 of the Yankees" in FIG. 6) or objects (e.g. the stadium name)

<u>Detailed Description Text</u> (33):
In this section, we present the explicit types of relationships between content elements that we propose to be included in the Generic AV DS. As shown in FIG. 4, relationships are defined at the different levels of the visual <u>structure</u> presented earlier. To represent relationships among content elements, we consider the division into syntax and <u>semantics</u> in the visual <u>structure</u>. Some of the limits among the relationship types that we propose are not rigid, as for the level of the visual <u>structure</u> discussed below.

<u>Detailed Description Text</u> (34):
Relationships at the syntactic levels of the visual <u>structure</u> can only occur in 2D space because there is no knowledge of objects at these levels to determine 3D relationships. At the syntactic levels, there can only be syntactic relationships, i.e. spatial (e.g. "Next to"), temporal (e.g. "In parallel"), and visual (e.g. "Darker than") relationships, which are based uniquely based on syntactic knowledge. Spatial and temporal attributes are classified in topological and directional classes. Visual relationships can be further indexed into global, local, and composition.

<u>Detailed Description Text</u> (35):
At the <u>semantic</u> levels of the visual <u>structure,</u> relationships among content elements could occur in 3D. As shown in FIG. 4, elements within these levels could be associated with not only <u>semantic</u> relationships but also syntactic relationships (e.g. "One person is next to another person", and "One person is a friend of another person"). We distinguish between two different types of <u>semantic</u> relationships: lexical relationships such as synonymy, antonymy,

hyponymy/hypernymy, and meronymy/holonymy; and predicative relationships referring to actions (events) or states.

Detailed Description Text (36):
In FIG. 4, relationships are proposed at different levels of the visual structure. Elements within the syntactic levels are related according to one types of relationships: syntactic. Elements within the semantic levels are related according to two types of relationships: syntactic and semantics.

Detailed Description Text (37):
We shall explain more extensively the syntactic and the semantic relationships with examples in sections below. Tables 1 and 2 summarize the indexing structures for the relationships including examples.

Detailed Description Text (43):
In a similar way in which the elements of the visual structure have different levels (generic, specific, and abstract), these types of syntactic relationships (see Table 1) can be defined in a generic level ("Near") or a specific level ("0.5 feet from"). For example, operational relationships such "To be the union of", "To be he intersection of", and "To be the negation of" are topological, specific relationships either spatial or temporal (see Table 1).

Detailed Description Text (45):
Semantic Relationships

Detailed Description Text (46):
Semantic relationships can only occur among content elements at the semantic levels of the ten-level conceptual structure. We divide the semantic relationships into lexical semantic and predicative relationships. Table 2 summarizes the semantic relationships including examples.

Detailed Description Text (47):
The lexical semantic relationships correspond to the semantic relationships among nouns used in WordNet. These relationships are synonymy (pipe is similar to tube), antonymy (happy is opposite to sad), hyponymy (a dog is an animal), hypernymy (an animal and a dog), meronymy (a musician is member of a musical band), and holonymy (a musical band is composed of musicians).

Detailed Description Text (48):
The predicative semantic attributes refer to actions (events) or states among two ore more elements. Examples of action relationships are "To throw" and "To hit". Examples of state relationships are "To belong" and "To own". FIG. 8 includes two action relationships: "Throw" and "Hit". Instead of only dividing the predicative semantic into actions or states, we could use the partial relational semantic decomposition used in WordNet. WordNet divides verbs into fifteen (15) semantic domains: verbs of bodily care and functions, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social interaction, and weather verbs. Only those domains that are relevant for the description of visual concept could be used.

Detailed Description Text (49):
As for the ten-level visual structure presented herein, we can define semantic relationships at different levels: generic, specific, and abstract. For example, a generic action relationship is "To own stock", a specific action relationship is "To own 80% of the stock", and, finally, an abstract semantic relationships is "To control the company".

Detailed Description Text (50):
For the Throwing and the Hitting Events in FIG. 6, FIG. 8 shows the use of semantic relationships to describe the actions of two objects: the Pitcher Object "Throws"

the Ball Object at the Batter Object and the Batter Object "Hits" the Ball Object.

Detailed Description Text (53):
We define types of content elements based on (1) the attributes that describe them
and (2) the relationships that associate them to other content elements.
Previously, we indexed the visual attributes of the content elements in a ten-level
visual structure. The first four levels of the pyramid correspond to syntax, and
the other six levels to semantics. Further, we divided the relationships into two
classes: syntactic and semantic. Consequently, we propose two basic types of
content elements: syntactic and semantic elements (see FIG. 5). Syntactic elements
can have only syntactic attributes and relationships (e.g. a color histogram
attribute and spatial relationship "On top of"); semantic elements can have not
only semantic attributes and relationships but also syntactic attributes and
relationships (e.g. an object can be described by a color histogram and a semantic
annotation descriptors). Our approach differs from the current Generic AV DS in
that our semantic (or high-level) elements include syntactic and semantic
information solving the rigid separation of the syntactic and the semantic
structures.

Detailed Description Text (54):
As shown in FIG. 5, we further classify the syntactic elements into region,
animated region, and segment elements. In a similar way, the semantic elements are
classified into the following semantic classes: object, animated object, and event.
Region and object are spatial entities. Segment and event are temporal entities.
Finally, animated-region and animated-object are hybrid spatial-temporal entities.
We explain each type in section accordingly.

Detailed Description Text (56):
The syntactic element is a content element in image or video data that is described
only by syntactic attributes, i.e. type, global distribution, local structure, or
global composition attributes (see FIG. 5). Syntactic elements can only be related
to other elements by visual relationships. We further categorize the syntactic
elements into region, animated-region, and segment elements. These elements are
derived from the syntactic element through inheritance relationships.

Detailed Description Text (58):
The segment element is a pure temporal entity that refers to an arbitrary set of
contiguous or not contiguous frames of a video sequence. A segment is defined by a
set of syntactic features, and a graph of segments, animated regions, and regions
that are related by temporal and visual relationships (see FIG. 5). The composition
relation is of type temporal, topological. Possible attributes of segments are
camera motion, and the syntactic features. For example, the Batting Segment in FIG.
7 is a segment element that is described by a temporal duration (global
distribution, syntactic), and shape mask (local structure, syntactic) attributes.
This segment has a "Consist of" relationship with the Batting Event (spatial-
temporal relationship, syntactic).

Detailed Description Text (59):
The animated-region element is a hybrid spatial-temporal entity that refers to an
arbitrary section of an arbitrary set frames of a video sequence. An animated
region is defined by a set of syntactic features, a graph of animated regions and
regions that are related by composition, spatial-temporal relationships, and visual
relationships (see FIG. 5). Animated regions may contain any features from the
region and the segment element. The animated region is a segment and a region at
the same time. For example, the Pitcher Region in FIG. 8 is an animated region that
is described by an aspect ratio (global distribution, syntactic), a shape mask
(local structure, syntactic), and a symmetry (global composition, syntactic)
attributes. This animated region is "On top of" the Sand 3 Region (spatial-temporal
relationship, syntactic).

Detailed Description Text (60):
Semantic Entities

Detailed Description Text (61):
The semantic element is a content element that is described by not only semantic
features but also by syntactic features. Semantic elements can be related to other
elements by semantic and visual relationships (see FIG. 5). Therefore, we derive
the semantic element from the syntactic element using inheritance. We further
categorize the semantic elements into object, animated-object, and event elements.
Pure semantic attributes are annotations, which are usually in text format (e.g. 6-
W semantic annotations, free text annotations).

Detailed Description Text (62):
The object element is a semantic and spatial entity; its refers to an arbitrary
section of an image or a frame of a video. An object is defined by a set of
syntactic and semantic features, and a graph of objects and regions that are
related by spatial (composition is a spatial relationship), visual, and semantic
relationships (see FIG. 5). The object is a region.

Detailed Description Text (63):
The event element is a semantic and temporal entity; its refers to an arbitrary
section of a video sequence. An event is defined by a set of syntactic and semantic
features, and a graph of events, segments, animated regions, animated objects,
regions, and objects that are related by temporal (composition is a temporal
relationship), visual, and semantic relationships. The event is a segment with
semantic attributes and relationships. For example, the Batting Event in FIG. 7 is
an event element that is described by a "Batting" (generic scene, semantic), "Bat
by player #32, Yankees" (specific scene, semantic), and a "Good Strategy" (abstract
scene, semantic) attributes. The syntactic attributes of the Batting Segment can
apply to the Batting Event (i.e. we could have not distinguished between Batting
Event and Batting Segment, and could have assigned the syntactic attributes of the
Batting Segment to the Batting Event). The Batting Event is composed of the Field
Object, and the Throwing and the Hitting Events, which represent the two main
actions in the Batting Event (i.e. throwing and hitting the ball). The Throwing and
the Hitting Events are related by a "Before" relationship (temporal relationship,
syntactic).

Detailed Description Text (64):
Finally, the animated-object element is a semantic and spatial-temporal entity; it
refers to an arbitrary section in an arbitrary set of frames of a video sequence.
An animated object is defined by a set of syntactic and semantic features, and a
graph of animated objects animated regions, regions, and objects that are related
by composition, spatial-temporal, visual, and semantic relationships (see FIG. 5).
The animated object is an event and an object at the same time. For example, the
Pitcher Object in FIG. 8 is an animated object that is described by "Man" (generic
object, semantic), "Player #3, Yankees" (specific object, semantic), and a
"Speed" (abstract object, semantic) attributes. This animated object is "On top of"
the Sand 3 Region shown in FIG. 9 (spatial-temporal relationship, syntactic). The
syntactic features of Pitcher Regions may apply to the Pitcher Object. We separate
the syntactic and the semantic attributes of this animated object as specified in
the Generic AV DS. However, we lose flexibility and efficiency in doing so because
we distribute the definition of the "real" object across different elements.

Detailed Description Text (65):
FIG. 5 provides fundamental models of each proposed type of content element.
Attributes, elements, and relationships are categorized in the following classes:
syntactic and semantic. The semantic and syntactic attributes have an associated
attribute, level, whose value correspond to the level of the visual that they refer
to. Syntactic elements are further divided in region, segment, and animated
regions. Semantic elements are categorized in object, animated object, and event

classes.

Detailed Description Text (67):
FIG. 7 provides a conceptual description of the Batting Event for the Baseball game
in FIG. 6 in accordance with the present invention.

Detailed Description Text (68):
FIG. 8 provides a conceptual description of the Hit and the Throw Events for the
Batting Event in FIG. 6 in accordance with the present invention.

Detailed Description Text (69):
FIG. 9 provides a conceptual description of the Field Object for the Batting Event
in FIG. 6 in accordance with the present invention.

Detailed Description Text (72):
One of the difficulties inherent in the indexing of images is the number of ways in
which they can be analyzed. A single image may represent many things, not only
because it contains a lot of information, but because what we see in the image can
be mapped to a large number of abstract concepts. A distinction between those
possible abstract descriptions and more concrete descriptions based only on the
visual aspects of the image constitutes an important step in indexing.

Detailed Description Text (73):
In the following sections, we make distinctions between percept and concept. We
then provide definitions for syntax and semantics, and finally discuss general
concept space and visual concept space. The importance of these definitions in the
context of content-based retrieval will be apparent when we define our indexing
structures.

Detailed Description Text (78):
Syntax and Semantics

Detailed Description Text (79):
In a similar way in which percepts require no interpretation, syntax refers to the
way visual elements are arranged without considering the meaning of such
arrangements. Semantics, on the other hand, deals with the meaning of those
elements and of their arrangements. As will be shown in the discussion that
follows, syntax can refer to several perceptual levels--from simple global color
and texture to local geometric forms such as lines and circles. Semantics can also
be treated at different levels.

Detailed Description Text (83):
These definitions are useful since they point out a very important issue in
content-based retrieval: different users have different concepts (of even simple
objects), and even simple objects can be seen at different conceptual levels.
Specifically, there is an important distinction between general concept (i.e.,
helps answer the question: what is it?) and visual concept (i.e., helps answer the
question: what does it look like?) and this must be considered when designing an
image database. We apply these ideas to the construction of our indexing
structures. Conceptual category structure may be based on perceptual structure.

Detailed Description Text (85):
As noted in the previous section, there are many levels of information present in
images, and their multi-dimensionality must be taken into account when organizing
them in a digital library. The first step in creating a conceptual indexing
structure is to make a distinction between visual and non-visual content. The
visual content of an image corresponds to what is direclty perceived when the image
is observed (i.e., descriptors stimulated directly by the visual content of the
image or video in question--the lines, shapes, colors, objects, etc). The non-
visual content corresponds to information that is closely related to the image, but

that is not explicitly given by its appearance. In a painting, for example, the price, current owner, etc. belong to the non-visual category. Next we present an indexing structure for the visual content of the image and we follow with a structure for non-visual information.

Detailed Description Text (88):
Our visual structure contains ten levels: the first four refer to syntax, and the remaining six refer to semantics. In addition, levels one to four are directly related to percept, and levels five through ten to visual concept. While some of these divisions may not be strict, they should be considered because they have a direct impact in understanding what the user is searching for and how he tries to find it in a database. They also emphasize the limitations of different indexing techniques (manual and automatic) in terms of the knowledge required. An overview of the structure is given in FIG. 3. Observing this figure from top to bottom, it is clear that at the lower levels of the pyramid, more knowledge and information is required to perform indexing. The width of each level gives an indication of the amount of knowledge required there--for example, more information is needed to name specific objects in a scene. Each level is explained below and a discussion of the relationship between levels appears thereafter.

Detailed Description Text (89):
Observing this structure, it will be apparent that most of the efforts in content-based retrieval have focused on syntax (i.e., levels one through four). Techniques to perform semantic classification at levels five through ten, however, are highly desirable. The structure we present, helps identify the level of attributes handled by a specific technique, or provided by a given description (e.g., MPEG-7 annotations).

Detailed Description Text (91):
At the most basic level, we are interested in the general visual characteristics of the image or the video sequence. Descriptions of the type of image or video sequence or the technique used to produce it are very general, but prove to be of great importance. Images, for example, may be placed in categories such as painting, black and white (b&w), color photograph, and drawing. Related classification schemes at this level have been done conceptually, and automatically in WebSEEk.

Detailed Description Text (92):
In the case of digital photographs, the two main categories could be color and grayscale, with additional categories/descriptions which affect general visual characteristics. These could include number of colors, compression scheme, resolution, etc. We note that some of these may have some overlap with the non-visual indexing aspects described herein.

Detailed Description Text (96):
Local Structure

Detailed Description Text (97):
In contrast to Global Structure, which does not provide any information about the individual parts of the image or the video sequence, the Local Structure level is concerned with the extraction and characterization of the image's components. At the most basic level, those components result from low-level processing and include elements such as the Dot, Line, Tone, Color, and Texture. In the Visual Literacy literature, some of these are referred to as the "basic elements" of visual communication and are regarded as the basic syntax symbols. Other examples of local structure attributes are temporal/spatial position (e.g. start time and centroid), local color (e.g. M.times.N Layout), local motion, local deformation, and local shape/2D geometry (e.g. bounding box). There are various images in which attributes of this type may be of importance. In x-rays and microscopic images there is often a strong concern for local details. Such elements have also been used in content-

based retrieval systems, mainly on query by user-sketch interfaces. The concern here is not with objects, but rather with the basic elements that represent them and with combinations of such elements--a square, for example, is formed by four lines. In that sense, we can include here some "basic shapes" such as circle, ellipse and polygon. Note that this can be considered a very basic level of "grouping" as performed by humans when perceiving visual information.

Detailed Description Text (99):
At this level, we are interested in the specific arrangement of the basic elements given by the local structure, but the focus is on the Global Composition. In other words, we analyze the image as a whole, but use the basic elements described above (line, circle, etc.) for the analysis.

Detailed Description Text (102):
Up to the previous level the emphasis had been on the perceptual aspects of the image. No world knowledge is required to perform indexing at any of the levels above, and automatic techniques rely only on low-level processing. While this is an advantage for automatic indexing and classification, studies have demonstrated that humans mainly use higher level attributes to describe, classify and search for images. Objects are of particular interest, but they can also be placed in categories at different levels--an apple can be classified as a Macintosh apple, as an apple or as a fruit. When referring to Generic Objects, we are interested in the basic level categories: the most general level of object description. In the study of art, this level corresponds to pre-Iconography, and in information sciences one refers to it as the generic of level. The common underlying idea in these concepts and our definition of Generic Objects is that only general everyday knowledge is necessary to recognize the objects. A Machintosh apple, for example, would be classified as an apple at this level: that is the most general level of description of that object.

Detailed Description Text (116):
We have chosen a pyramid representation because it directly reflects several important issues inherent in our structure. It is apparent that at the lower levels of the pyramid, more knowledge and information is required to perform the indexing. This knowledge is represented by the width of each level. It is important to point out, however, that this assumption may have some exceptions. An average observer, for example, may not be able to determine the technique that was used to produce a painting--but an expert in art would be able to determine exactly what was used. Indexing in this particular case would require more knowledge at the type/technique level than at the generic objects level (since special knowledge about art techniques would be needed). In most cases, however, the knowledge required for indexing will increase in our structure from top to bottom: more knowledge is necessary to recognize a specific scene (e.g., Central Park in New York City) than to determine the generic scene level (e.g., park).

Detailed Description Text (119):
In this section, we briefly present a representation for relations between image elements.sup.8. This structure accommodates relations at different levels and is based on the visual structure presented earlier. We note that relations at some levels are most useful when applied between entities to which the structure is applied (e.g., scenes from different images may be compared). Elements within each level are related according to two types of relations: syntactic and semantic (only for levels 5 through 10). For example: two circles (local structure) can be related spatially (e.g., next to), temporally (e.g., before) and/or visually (e.g., darker than). Elements at the semantic levels (e.g., objects) can have syntactic and semantic relations--(e.g., two people are next to each other, and they are friends). In addition, each relation can be described at different levels (generic, specific, and abstract). We note that relations between levels 1, 6, 8, and 10 can be most useful between entities represented by the structure (e.g., between images, between parts of images, scenes, etc.)

Detailed Description Text (120):
The visual structure may be divided into syntax/percept and visual
concept/semantics. To represent relations, we observe such division and take into
consideration the following: (1) Knowledge of an object embodies knowledge of the
object's spatial dimensions, that is, of the gradable characteristics of its
typical, possible or actual, extension in space; (2) knowledge of space implies the
availability of some system of axes which determine the designation of certain
dimensions of, and distances, between objects in space. We use this to argue that
relations that take place in the syntactic levels of the visual structure can only
occur in 2D space since no knowledge of the objects exist (i.e., relationships in
3D space cannot be determined). At the local structure level, for example, only the
basic elements of visual literacy are considered, so relations at that level are
only described between such elements (i.e., which do not include 3D information).
Relations between elements of levels 5 through 10, however, can be described in
terms of 2D or 3D.

Detailed Description Text (121):
In a similar way, the relations themselves are divided into the classes syntactic
(i.e., related to perception) and semantic (i.e. related to meaning). Syntactic
relations can occur between elements at any of the levels, but semantic relations
occur only between elements of levels 5 through 10. Semantic relationships between
different colors in a painting, for example, could be determined (e.g., the
combination of colors is warm), but we do not include these at that level of our
model.

Detailed Description Text (123):
Temporal relations refer to those that connect elements with respect to time (e.g.,
in video these include before, after, between, etc.), and visual relations refer
only to visual features (e.g., bluer, darker, etc.). Semantic relations are
associated with meaning (e.g., owner of, friend of, etc.).

Detailed Description Text (124):
In a similar way in which the elements of the visual structure have different
levels (generic, specific, abstract), relations can be defined at different levels.
Syntactic relations can be generic (e.g., near) or specific (e.g, a numerical
distance measure). Semantic relationships can be generic, specific, or abstract.

Detailed Description Text (125):
As an example, spatial global distribution could be represented by a distance
histogram, local structure by relations between local components (e.g., distance
between visual literacy elements), and global composition by global relations
between visual literacy elements.

Detailed Description Text (127):
As explained at the beginning of this section, non-visual information refers to
information that is not directly part of the image, but is rather associated with
it in some way. One may divide attributes into biographical and relationship
attributes. While it is possible for non-visual information to consist of sound,
text, hyperlinked text, etc., our goal here is to present a simple structure that
gives general guidelines for indexing. We will focus briefly on text information
only. FIG. 10 gives an overview of this structure.

Detailed Description Text (131):
The second class of non-visual information is directly linked to the image in some
way. Associated Information may include a caption, article, a sound recording, etc.
As discussed, in many cases this information helps perform some of the indexing in
the visual structure, since it may contain specific information about what is
depicted in the image (i.e., the subject). In that context, it is usually very
helpful at the semantic levels since they require more knowledge that is often not

present in the image alone. In some cases, however, the information is not directly related to the subject of the image, but it is associated to the image in some way. A sound recording accompanying a portrait, for example, may include sounds that have nothing to do with the person being depicted--they are associated with the image though, and could be indexed if desired.

Detailed Description Text (134):
Relationships Between Indexing Structures

Detailed Description Text (135):
We define a Semantic Information Table to gather high level information about the image (ee FIG. 11). The table can be used for individual objects, groups of objects, the entire scene, or parts of the image. In most cases visual and non-visual information contribute in filling in the table--simple scene classes such as indoor/outdoor may not be easily determined from the visual content alone; location may not be apparent from the image, etc. Individual objects can be classified and named based on the non-visual information, contributing to the mapping between visual object and conceptual object.

Detailed Description Text (136):
In FIG. 11, visual and non-visual information can be used to semantically characterize an image or its parts. The way in which these two modalities contribute to answer the questions in the semantic table may vary depending on the content. The table helps answer questions such as: What is the subject (person/object, etc.)?, What is the subject doing? Where is the subject? When? How? Why? The table can be applied to individual objects, groups of objects, the entire scene, or parts of the image.

Detailed Description Text (137):
The relationship between this structure and the visual structure is apparent when applying the table at each level beginning with level 5. We also note that while the table provides a compact representation for some information related to the image, it does not replace the indexing structures presented. The group of structures provides the most complete description.

Detailed Description Text (138):
Having the appropriate indexing structures, we can focus on how the contents of a digital library may be organized. In the next section, we analyze issues that play a crucial role in the organization and retrieval of images.

Detailed Description Text (140):
In order to be successful at building an image digital library, it is not only important to understand the data, but also the human issues related to classification. In this section we discuss issues of importance in this respect, and explain how we apply the concepts in building our image indexing test bed. First, we discuss categories. Then, we discuss levels and structure in categorization. Finally, we present some of the issues related to attributes and similarity.

Detailed Description Text (143):
In our structure we can identify Sensory Perception categories such as color and texture. GK categories, however, play a very important role since users are mainly interested in the objects that appear in the images and what those objects may represent. Some theories in cognitive psychology express that classification in GK categories is done as follows:

Detailed Description Text (148):
Category Structure

Detailed Description Text (149):

Category <u>structure</u> is a crucial factor in a digital library and brings about several issues of importance which we briefly discuss here. The following issues should be considered: relationships between categories (e.g., hierarchical or entity-relation), the levels of abstraction at which classification should be performed (e.g., studies by Rosch) suggest the existence of a basic level and subordinate/superordinate level categories), horizontal category <u>structure</u> (i.e., how each category should be organized and the degrees of membership of elements within each category--these can be <u>fuzzy</u> or binary), etc.

<u>Detailed Description Text</u> (150):
In addition to considering different levels of analysis when indexing visual information, the way in which similarity is measured is of great importance. Issues related to measurements of similarity include the level of consideration (e.g., part vs. whole), the attributes examined, the types of attributes (e.g., levels of our <u>structures</u>), whether the dimensions are separable or not, etc.

<u>Detailed Description Text</u> (152):
We are developing an image indexing test bed that incorporates the concepts presented herein, using different techniques to index images based on the <u>structure</u> set forth herein. In particular, for type/technique we are using discriminant analysis. For global distribution, we use global color histograms and Tamura texture measures. At the local <u>structure</u> level, we allow sketch queries as in VideoQ, by using automatic segmentation and also multi-scale phase-curvature histograms of coherent edge-maps and projection histograms. Global composition is obtained by performing automatic segmentation and merging of generated regions to yield iconic representations of the images.

<u>Detailed Description Text</u> (157):
Another illustrative discussion of the advantages of the present invention may be provided by setting forth an exemplary <u>description</u> of its use in conjunction with a digital signal that represents audio content.

<u>Detailed Description Text</u> (158):
We previously proposed a ten-level conceptual <u>structure</u> to index the visual content elements (e.g. regions, entire images, events, etc.) of images. The classification in that work refers only to descriptors for visual content (i.e., not meant for "metadata"--for example, the name of the person who took the photograph is not a visual descriptor).

<u>Detailed Description Text</u> (159):
In this document, we propose the classification of audio descriptors (to be included in the <u>MPEG-7</u> audio part of the standard), based on the ten-level conceptual <u>structure</u> presented earlier. The pyramid <u>structure</u> we propose contains exactly the same levels as the visual <u>structure</u> previously described in connection with FIG. 3 and FIG. 4. Each level, however, refers to audio elements instead of visual elements. In the original <u>structure,</u> an object corresponds to a visual entity. In the new <u>structure,</u> an object corresponds to an audio entity (e.g., a person's voice).

<u>Detailed Description Text</u> (160):
The importance of the separation between syntax and <u>semantics</u> has been widely identified by researchers in the area of image and video indexing. Although we are not aware of similar studies for audio content, the results from the studies examined suggest that this separation is very useful in audio indexing also. For instance, studies in information retrieval and cognitive psychology have shown how individuals use different levels to describe (or index) images/objects. While some of the divisions we present may not be strict, they should be considered because they have a direct impact on how the audio content is indexed, handled and presented to the users (e.g., applications or human viewers) of such content.

Detailed Description Text (161):
The structure presented earlier for visual attributes, which draws on research from
different fields related to image indexing, provides a compact and organized
classification that can be easily applied to audio. The structures are intuitive
and highly functional and stress the need, requirements, and limitations of
different indexing techniques (manual and automatic). The indexing cost
(computational or in terms of human effort) for an audio segment, for example, is
generally higher at the lower levels of the pyramid: automatically determining the
type of content (music vs. voice) vs. recognizing generic objects (e.g., voice of a
man) vs. recognizing specific objects (e.g., voice of Bill Clinton). This also
implies that more information/knowledge is required at the lower levels and if a
user (e.g. application) makes a request to another user (e.g., application), there
will be clarity regarding how much additional information might be needed, or what
level of "service" a user can expect from, say, a level 5 audio classifier. In
addition, this breakdown of the attributes and relationships is of great value
since humans often make comparisons based on attributes. The benefits of the
structures proposed have been shown in preliminary experiments for visual content,
and efforts to conduct core experiments are also being made. These experiments, and
the flexibility that allows the use of the structure for audio indexing suggest the
benefits of applying this sort of descriptor classification to audio and visual
content.

Detailed Description Text (164):
The proposed audio structure contains ten levels: the first four refer to syntax,
and the remaining six refer to semantics. An overview for the audio structure can
be drawn from FIG. 3. The width of each level in an indication of the amount of
knowledge/information required. The syntax levels are type/technique, global
distribution, local structure, and global composition. The semantic levels are
generic object, generic scene, specific object, specific scene, abstract object,
and abstract scene.

Detailed Description Text (165):
The syntax levels classify syntactic descriptors, that is, those that describe the
content in terms of low-level features. In the visual structure, these referred to
the colors and textures present in the image. In the audio structure of this
document, they refer to the low-level features of the audio signal (whether it is
music, voice, etc.). Examples include the fundamental frequency, harmonic peaks,
etc.

Detailed Description Text (166):
The semantic levels of the visual structure classified attributes related to
objects and scenes. The semantic levels in the audio structure are analogous,
except that the classification is based on the attributes extracted from the audio
signal itself. Like in the visual case, in audio it is possible to identify objects
(e.g., voice of a man, sound of a trumpet, etc.), and scenes (e.g., street noise,
opera, etc.).

Detailed Description Text (167):
Each level of the visual structure, which is analogous, has been explained
previously. Next, we briefly explain each level and describe how it can be used for
the classification of audio descriptors. We use the words attribute and descriptor
interchangeably, and give intuitive examples for each level, making analogies with
the visual structure to help clarify the explanations. For the semantic levels, it
is useful to think of a typical radio news broadcast, in which different entities
are used interchangeably--persons, noises, music, and scenes (e.g., it is common in
on-site reports to hear background noise or music, during, before and after a
journalist's report).

Detailed Description Text (169):
General descriptions of the type of audio sequence. For example: music, noise,

voice, or any combination of them: stereo, number of channels, etc.

Detailed Description Text (171):
Attributes that describe the global content of audio, measured in terms of low-level features. The attributes at this level are global because they are not concerned with individual components of the signal, but rather with a global description. For example, a signal can be described as being Gaussian noise--such description is global because it doesn't say anything about the local components (e.g., what elements, or low-level features describe the noise signal).

Detailed Description Text (172):
Local Structure

Detailed Description Text (173):
Concerned with the extraction and characterization of individual low-level syntactic components in the audio segment. In contrast to the previous level, attributes here are meant to describe the local structure of the signal. In an image, the local elements are given by basic syntax symbols that are present in the image (e.g., lines, circles, etc.). This level serves the same function in audio, so any low-level (i.e., not semantic such as a word, or a letter in spoken content) local descriptor would be classified at this level.

Detailed Description Text (175):
Global description of an audio segment based on the specific arrangement or composition of basic elements (i.e., the local structure descriptors). While local structure focuses on specific local features of the audio, Global Composition focuses on the structure of the local elements (i.e., how they are arranged). For example, an audio sequence can be represented (or modeled) by a Markov chain, or by any other structure that uses low-level local features.

Detailed Description Text (177):
Up to the previous level, no world knowledge is required to perform indexing-quantitative features can be automatically extracted from the audio segment and classified into the syntactic levels described. When the audio segment is described in terms of semantics (e.g., recognition), however, objects play an important role. Objects, however, can be placed in categories at different levels--an apple can be classified as a Macintosh apple, as an apple, or as a fruit. The recognition of an object, can be based on an audio segment, and therefore we can make a similar classification. For example, we can say that an audio entity corresponds (e.g., a voice) to a man, or to Bill Clinton. When referring to Generic Objects, we are interested in the basic level categories: the most general level of object description, which can be recognized with everyday knowledge. That means there is no knowledge of the specific identity of the object in question (e.g., explosion, rain, clap, man's voice, woman's voice, etc.). Audio entity descriptors can be classified at this level.

Detailed Description Text (187):
The Abstract Scene level refers to what the audio segment as a whole represents. It may be very subjective. For images, it has been shown, for example, that users sometimes describe images in affective (e.g. emotion) or abstract (e.g. atmosphere, theme) terms. Similar descriptions can be assigned to audio segments, for example, attributes to describe an audio scene could include: sadness (e.g., people crying), happiness (e.g., people laughing), etc.

Detailed Description Text (190):
In this section, we present the explicit types of relationships between content elements that we propose. These relationships are analogous to those presented earlier for visual content. As shown in FIG. 12, relationships are defined at the different levels of the audio structure presented earlier in connection with FIG. 3. To represent relationships among content elements, we consider the division into

syntax and <u>semantics</u>.

<u>Detailed Description Text</u> (191):
At the syntactic levels, there can be syntactic relationships, i.e. spatial (e.g.,
"sound A is near sound B"), temporal (e.g. "In parallel"), and audio (e.g. "Louder
than") relationships, which are based uniquely on syntactic Knowledge. Spatial and
temporal attributes are classified in topological and directional classes. Audio
relationships can be further indexed into global, local, and composition. As shown
in FIG. 12, elements within these levels could be associated with not only <u>semantic</u>
relationships, but also syntactic relationships (e.g. "the trumpet sounds near the
violin", and "the trumpet notes complement the violin notes"). We distinguish
between two different types of <u>semantic</u> relationships: lexical relationships such
as synonymy, antonymy, hyponymy/hypernymy, and meronymy/holonymy; and predicative
relationships referring to actions (events) or states.

<u>Detailed Description Text</u> (193):
We shall explain more extensively the syntactic and the <u>semantic</u> relationships with
examples. Tables 3 and 4 below summarize the indexing <u>structures</u> for the
relationships including examples.

<u>Detailed Description Text</u> (198):
Audio relationships relate audio entities based on their visual attributes or
features. These relationships can be indexed into global, local, and composition
classes (see Table 3). For example, an audio global relationship could be "To be
less noisy than" (based on a global noise feature), an audio local relationship
could be "is louder than" (based on a local loudness measure), and an audio
composition relationship could be based on comparing the <u>structures</u> of a Hidden
Markov Models.

<u>Detailed Description Text</u> (199):
In a similar way in which the elements of the audio <u>structure</u> have different levels
(generic, specific, and abstract), these types of syntactic relationships (see
Table 3) can be defined in a generic level ("Near") or a specific level ("10 meters
from"). For example, operational relationships such "To be the union of", "To be
the intersection of", and "To be the negation of" are topological, specific
relationships either spatial or temporal (see Table 3).

<u>Detailed Description Text</u> (200):
<u>Semantic</u> Relationships

<u>Detailed Description Text</u> (201):
<u>Semantic</u> relationships can only occur among content elements at the <u>semantic</u> levels
of the ten-level conceptual <u>structure</u>. We divide the <u>semantic</u> relationships into
lexical and predicative relationships. Table 4 summarizes the <u>semantic</u>
relationships including examples. Note that since <u>semantic</u> relationships are based
on understanding of the content, we can make the same classification for
relationships obtained from visual content as for relationships obtained from audio
content. The <u>semantic</u> relationships here, therefore, are identical to those
described in connection with video signals. The only difference lies in the way the
<u>semantic</u> content is extracted (i.e., understanding the audio vs. understanding an
image or video). To make the explanation more clear, we have used examples related
to audio, although the original examples would also apply. For instance: that apple
is like that orange, as a generic synonymy example--the apple and orange could be
"recognized" from the audio, if a speaker talks about them.

<u>Detailed Description Text</u> (202):
The lexical <u>semantic</u> relationships correspond to the <u>semantic</u> relationships among
nouns used in WordNet. These relationships are synonymy (violin is similar to a
viola), antonymy (flute is opposite to drums), hyponymy (a guitar is a string
instrument), hypernymy (a string instrument and a guitar), meronymy (a musician is

member of a musical band), and holonymy (a musical band is composed of musicians).

Detailed Description Text (203):
The predicative semantic attributes refer to actions (events) or states among two
ore more elements. Examples of action relationships are "To yell at" and "To
hit" (e.g., hit a ball). Examples of state relationships are "To belong" and "To
own". Instead of only dividing the predicative semantic into actions or states, we
could use the partial relational semantic decomposition used in WordNet. WordNet
divides verbs into 15 semantic domains: verbs of bodily care and functions, change,
cognition, communication, competition, consumption, contact, creation, emotion,
motion, perception, possession, social interaction, and weather verbs. Only those
domains that are relevant for the description of visual concept could be used.

Detailed Description Text (204):
As for the ten-level audio structure presented herein, we can define semantic
relationships at different levels: generic, specific, and abstract. For example, a
generic action relationship is "To own stock", a specific action relationship is
"To own 80% of the stock", and, finally, an abstract semantic relationships is "To
control the company".

Detailed Description Text (205):
The present invention includes not only methods, but also computer-implemented
systems for multiple level classifications of digital signals (e.g., multimedia
signals) for indexing and/or classification purposes. The methods described
hereinabove have been described at a level of some generality in accordance with
the fact that they can be applied within any system for processing digital signals
of the type discussed herein--e.g., any of the art-recognized (or future-developed)
systems compatible with handling of digital multimedia signals or files under the
MPEG-7 standards.

Detailed Description Text (207):
To give a broad example, one could consider an exemplary embodiment of a system for
practicing the present invention in conjunction with any multimedia-compatible
device for processing, displaying, archiving, or transmitting digital signals
(including but not limited to video, audio, still image, and other digital signals
embodying human-perceptible content), such as a personal computer workstation
including a Pentium microprocessor, a memory (e.g., hard drive and random access
memory capacity), video display, and appropriate multimedia appurtenances.

Detailed Description Text (210):
We make the distinction between syntax and semantics for attributes (or MPEG-7
descriptors), relationships, and content elements. Syntax refers to the way the
content elements are arranged without considering the meaning of such arrangements.
Semantics, on the other hand, deals with the meaning of those elements and of their
arrangements. Syntactic and semantic attributes can refer to several levels.
Similarly, syntactic and semantic relationships can be further divided into sub-
types referring to different levels. We provide compact and clear definitions of
the syntactic and semantic elements based on their types of attributes and
relationships with other elements. An importance difference with the Generic AV DS,
however, is that our semantic elements include not only semantic attributes but
also syntactic attributes. Therefore, if an application would rather not
distinguish between syntactic and semantic elements, it can do so by using only
semantic elements.

Detailed Description Paragraph Table (1):
TABLE 1 Indexing structure for syntactic relationships and examples. Types of
relationships Levels Examples Syntac- Spatial Topolog- Generic .sctn. Near from,
Far from, tic ical Adjacent to, Contained in, Composed of; Consist of
Specific .sctn. The union, The inter- section, The negation .sctn. 0.5 inches from,
the intersection of two regions .sctn. R in R.theta. Direc- Generic .sctn. Left of,

Top of, Upper tional left of, Lower right of. Behind .sctn. 2D String
Specific .sctn. The union, The inter- section, The negation .sctn. 20 degrees north
from, 40 degrees east from, the union of two segments .sctn. .theta. in R.theta.
Tempo- Topolog- Generic .sctn. Co-begin, Parallel, ral ical Sequential, Overlap,
Adjacent, Within, Composed, Consist of .sctn. SMIL's <seq> and <par>
Specific .sctn. 20 min. apart from, 20 sec. Overlapping .sctn. SMIL's <seq> and
<par> with attributes (start time, end time, duration) Direc- Generic .sctn.
Before, After tional Specific .sctn. 20 min. after Visual Global Generic .sctn.
Smother than, Darker than, More yellow than, Similar texture, Similar Color,
Similar speed Specific .sctn. Distance in texture feature, Distance in color
histogram .sctn. Indexing hierarchy based on color histogram Local Generic .sctn.
Faster than, To grow slower than, Similar speed, Similar shape Specific .sctn. 20
miles/hour faster than, Grow 4 inches/sec. faster than .sctn. Indexing hierarchy
based on local motion, deformation features Composi- Generic .sctn. More symmetric
than, tion Specific .sctn. Distance in symmetry feature .sctn. Indexing hierarchy
based on symmetry feature

Detailed Description Paragraph Table (2):
TABLE 2 Indexing structure for semantic relationships and examples. Types of
attributes Levels Examples Seman- Lexical Synonymy Generic That apple is like that
orange tic To be Similar to Specific That apple has as many calories as that orange
Abstract That apple is as nutritious as that orange Antonymy Generic That man is
different from that woman To be opposite to Specific That man is 20-pound heavier
than that woman Abstract That man is uglier than that woman Hypo- Generic A dog is
an animal nymy Hyper- nymy To be a - To be a type of Specific A dog is a mammal
animal Abstract A dog is a playful animal Mero- Generic Peter is a member of a team
nymy Holo- nymy To be a part/ member of - To be the whole of Specific Peter is an
outfielder for the Yankees Abstract Peter is the best outfielder in the Yankees'
history Predica- Action/ Generic The boys are playing with tive Event the girls
Specific The boys are playing soccer with the girls Abstract The boys are playing
soccer well with the girls State Generic The girl owns stock from that company
Specific The girl owns 80% of the stock from the company Abstract The girl controls
the company

Detailed Description Paragraph Table (3):
TABLE 3 Indexing structure for syntactic relationships and examples. Types of
relationships Levels Examples Syntac- Spatial Topolog- Generic .sctn. Near from,
Far from, tic ical Adjacent to, Contained in, Composed of; Consist of
Specific .sctn. The union, The inter- section, The negation, Normal decompositon,
Free decompositon .sctn. 10 meters from Direc- Generic .sctn. Left of, Top of,
Upper tional left of, Lower right of. Behind Specific .sctn. The union, The inter-
section, The negation .sctn. 20 degrees north from, 40 degrees east from, the union
of two segments Tempo- Topolog- Generic .sctn. Co-begin, Parallel, ral ical
Sequential, Overlap, Adjacent, Within, Composed, Consist of .sctn. SMIL's <seq> and
<par> Specific .sctn. 20 min. apart from, 20 sec. Overlapping .sctn. SMIL's <seq>
and <par> with attributes (start time, end time, duration) Direc- Generic .sctn.
Before, After tional Specific .sctn. 20 min. after Audio Global Generic .sctn.
Louder than, Softer than, Similar speed Specific .sctn. Distance in global
feature .sctn. Indexing hierarchy based on global feature Local Generic .sctn.
Faster than, To grow slower than, Similar speed, Similar shape Specific .sctn. 5 dB
louder than .sctn. Indexing hierarchy based on pitch Composi- Generic .sctn. HMM
structure A has the tion same number of states as structure B Specific .sctn.
Distance in composition feature .sctn. Indexing hierarchy based on composition
features

Detailed Description Paragraph Table (4):
TABLE 4 Indexing structure for semantic relationships and examples. Types of
attributes Levels Examples Seman- Lexical Synonymy Generic Symphony A is like tic
Symphony B To be Similar to Specific Symphony A has as many movements as Symphony B
Abstract Symphony A is as sad as Symphony B Antonymy Generic Sound A is different